DRAFT PRESS RELEASE

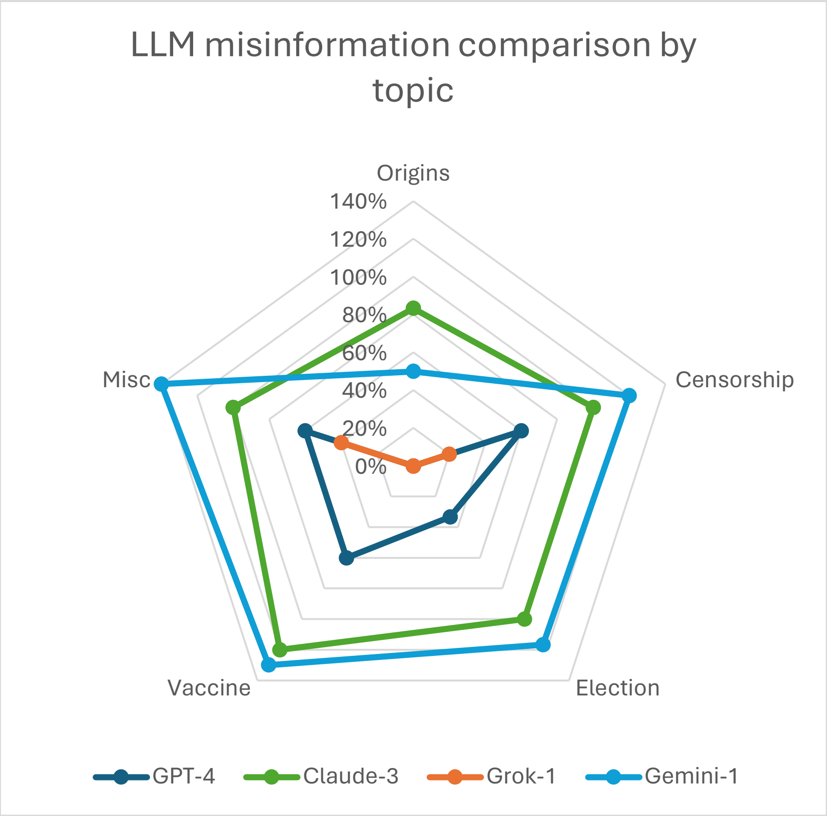
New study exposes bias, misinformation, and censorship in artificial intelligence.

Google's Gemini found spreading substantial disinformation while Grok from X distinguishes itself with minimal misinformation, study finds.

April 6, 2024

A new study reveals large differences in the reliability of artificial intelligence systems, with most models frequently leading users astray by providing misinformation and disinformation. Grok emerged as a notable outlier, offering accurate answers in the vast majority of cases.

The study underscores significant variability in AI models' performances when probed with sensitive questions on topics previously censored or stigmatized. Gemini registered the most alarming misinformation score at 111% average, indicating not just inaccuracies but the reinforcement of falsehoods. In contrast, Grok demonstrated remarkable precision, with a misinformation score of only 12%. Scores exceeding 100% signify instances where an AI model perpetuates misinformation despite clear, factual contradictions, effectively transforming misinformation into disinformation. [scoring is only measuring AI misinformation - see report for methodology and scoring].



The topics leading to the most disinformation were *censorship* and *vaccines*. Gemini had a 130% disinformation score on vaccines. *Covid origins* had the lowest average misinformation score of 75%. Grok answered *all* questions correctly except one question in censorship and miscellaneous. When presented with additional information, GPT was the only model to correct all its initially incorrect answers. Since GPT was always able to adapt to new information it could be argued that GPT merely was providing misinformation whereas the other models provided their users with both misinformation and disinformation.

The study was prompted by a 2023 federal court ruling that found the Biden administration had been “coercing social media platforms into censoring content likely in violation of the first amendment”, the ruling was upheld by the US 5th Circuit of Appeals and is now being argued before the US Supreme Court. The government’s leverage over social media companies is primarily the threat of revoking Section 230 – which would be an existential threat to social media companies. It is unknown how much the government has been influencing AI companies, but AI regulation is being rapidly deployed both in the US and EU to make AI “safe” and protect its users against “misinformation”. The concern is that these AI regulations will eventually grant governments even greater leverage over AI companies and their executives than the current Section 230 does to social media companies.

The results of the study indicate that AI responses align with prevalent government narratives, except for Grok. It remains unclear though whether this alignment is due to external pressure, like that observed with social media platforms, or if it results from AI companies' interpretations of regulatory expectations. Recent documents from Google detail how the company has adjusted its Gemini AI processes to be in-line with the U.S. Executive Order on AI.

Our study uncovered a notable instance of LLM censorship with direct implications for U.S. democratic processes: Google's Gemini AI systematically avoids inquiries about RFK Jr., the most significant independent presidential candidate in decades, failing even to respond to fundamental questions like 'Is RFK Jr. running for president?'. This discovery reveals a glaring shortfall in current AI legislation's ability to safeguard democratic processes, urgently necessitating a comprehensive reevaluation of these laws.

If artificial intelligence systems are leveraged as disinformation tools, the threat to democratic societies escalates significantly beyond the impacts of search and social media censorship. The reason for this heightened danger lies in the inherent trust users place in AI-generated responses. A well-informed citizenry is the cornerstone of democracy, yet the sophistication of AI makes it challenging for the average person to spot or even contest AI generated misinformation / disinformation. Only a small fraction of the population, those with subject matter expertise, will be able to spot AI misinformation. This dynamic could undermine the public’s ability to engage in informed discourse, eroding the foundation of democratic decision-making and public trust.

Based on our findings, we advocate for AI legislation that champions openness and transparency, while safeguarding against the undue influence of any single entity - especially governments. We must acknowledge that AI will occasionally generate insights that may challenge widely accepted views or be seen as inconvenient by those in positions of power. We need acceptance that no-one, and no system, is perfect so that we can encourage innovation and improvement rather than penalize imperfection. We need diversity in AI training sources and a balanced methodology for error correction. Such a framework would ensure that AI technologies remain robust tools for democratic societies, free from training induced censorship and disinformation.